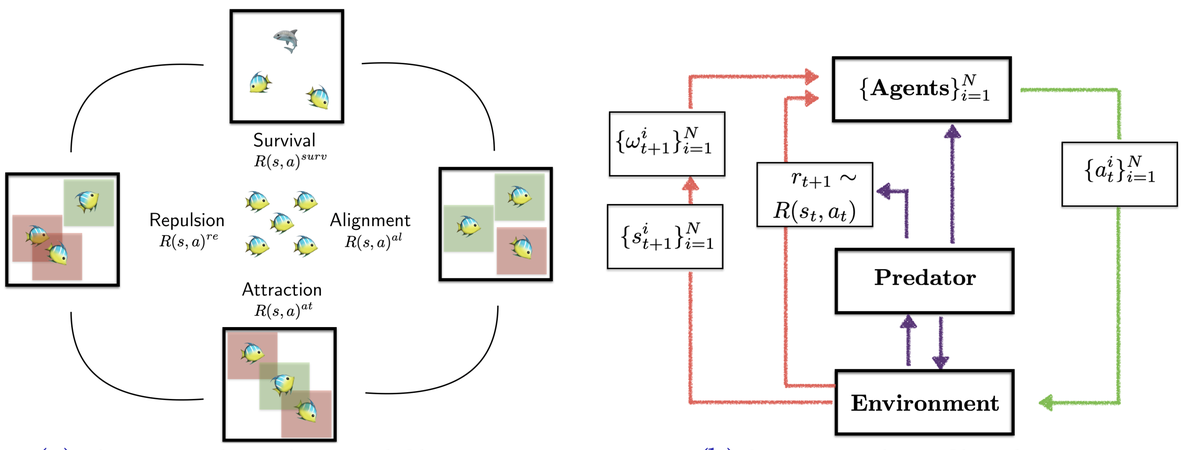
Deep Multi-Agent Reinforcement Learning for Complex Swarm Behavior

Robert Tjarko Lange (@RobertTLange)

Technical University Berlin (@SprekelerLab)
Einstein Center for Neurosciences Berlin & SCIoI Excellence Cluster

Motivation

- We introduce a normative approach for learning intelligent swarm behavior based on phenomenologically-grounded rewards and Deep MARL.
- While traditional work focuses on an evolutionary stochastic dynamical systems perspective (e.g. Romanczuk & Schimansky-Geier (2012)), little efforts have been made to train a large set of cooperative agents in a reward-based fashion.



(a) Phenomenological Reward Objectives (b) Swarm-Predator Closed Loop

- **Contributions**: We analyze the effect of local rewards and communication on swarm learning dynamics and introduce a reward curriculum in order to study competing behavioral traits.

Swarm Environment Design

- **Reward Design**: Inspired by phenomenological observations.
  - ⋆ *Attraction & Repulsion*: Agents desire to stay connected. The danger of collision, on the other hand, leads to distress and negative reinforcement.
  - ⋆ *Alignment*: Due to connectivity agents intend to move in the similar direction.
  - ⋆ *Survival*: Agents receive a negative reward once they collide with the predator.

$$R(\{s^{i'}\}|\{s^i\},\{a^i\}) = \frac{1}{N(N-1)}\left(R(\{s^i\},\{a^i\})^{at} + R(\{s^i\},\{a^i\})^{re} + R(\{a^i\})^{al} + R(\{s^i\},\{a^i\})^{surv}\right)$$

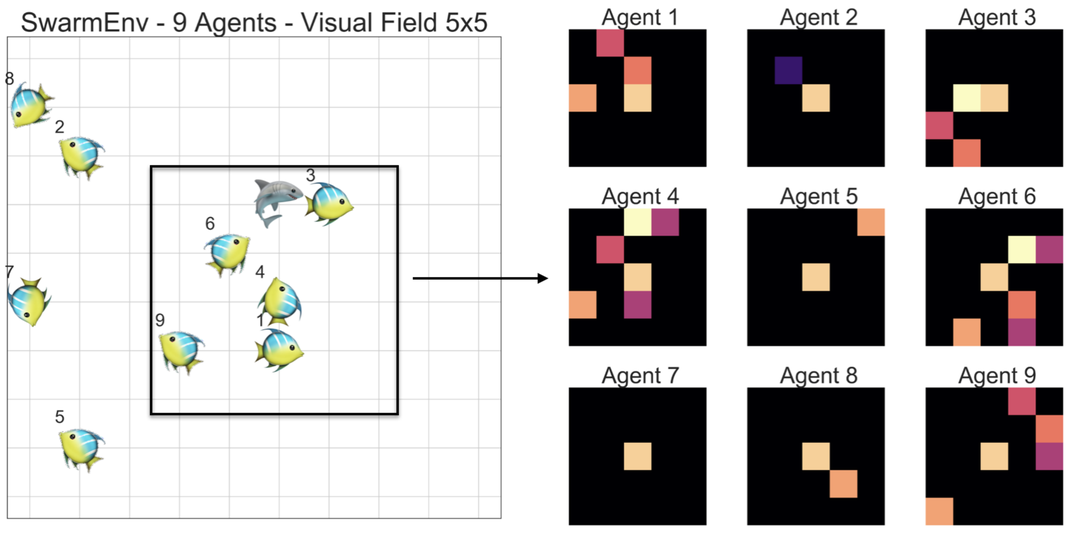


Figure 1: State Space and Observation Construction

- **Adversarial Predator**: Agent which follows the nearest target (updated ϵ-greedily). Once, they collide the episode terminates.
- **State and Action Space**: 2-D discrete state space with periodic boundary conditions. Instantaneous movement in all eight directions.
- **Partial Observability** - $\{o^i\}^N_{i=1} \sim \mathcal{O}(\{s^i\}^N_{i=1}; z)$: Each agent receives a receptive field of dimensionality $z \times z$. They can differentiate between the orientations of the cooperative agents as well as the adversarial predator.

Multi-Agent Reinforcement Learning

- **Decentralized Training via Independent DQN** (I-DQNs; Tampuu *et al.* (2017)): Given a set of N agents, each agent is associated with a shared network architecture and agent-specific parameters θ^i.
- The I-DQN agents are trained independently based on their own replay buffer. Mean Squared Bellman Error Objective:

$$\mathcal{L}^i_{I-DQN} = \mathbb{E}_{s^i,a^i,r,s^{i'}\sim\mathcal{D}^i}\left[\left(Q(s^i,a^i;\theta^i_k) - Y^i_k\right)^2\right]$$

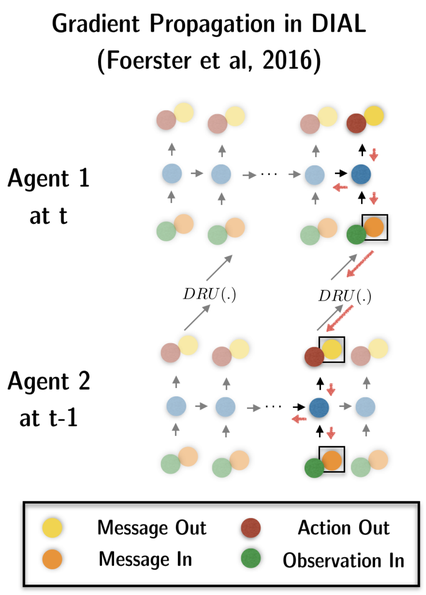
$$Y^i_k = r + \gamma \max_{a'\in\mathcal{A}} Q(s^{i'},a^{i'};\theta^{i|-}_k)$$

- **Centralized Training via DIAL** (Foerster *et al.* , 2016): Introduce differentiable communication and propagate gradient through agents.

Action: $Q_a(o^i_t, m^i_{t-1}, h^i_{t-1}, a^i_{t-1}, m^i_{t-1}, i, a^i_t)$

Messages: $Q_m(o^i_t, m^i_{t-1}, h^i_{t-1}, a^i_{t-1}, m^i_{t-1}, i, m^i_t)$

$$DRU(m^i_t) = \begin{cases} Logistic(\mathcal{N}(m^i_t, \sigma)), & \text{Training} \\ \mathbf{1}\{m^i_t > 0\}, & \text{Execution} \end{cases}$$



Learning from Local & Global Rewards

- The global reinforcement signal is an aggregation over the agent-specific reward. E.g. $R(\{a^i\})^{al} = -\sum^n_{j=1} r^{al}(a^j, a^{\neg j}) = -\sum^n_{j=1}\sum_{k\neq j}\frac{1-\cos(a^j,a^k)}{2}$. Instead of providing the same global reward to all agents, we can also reinforce each local contribution:

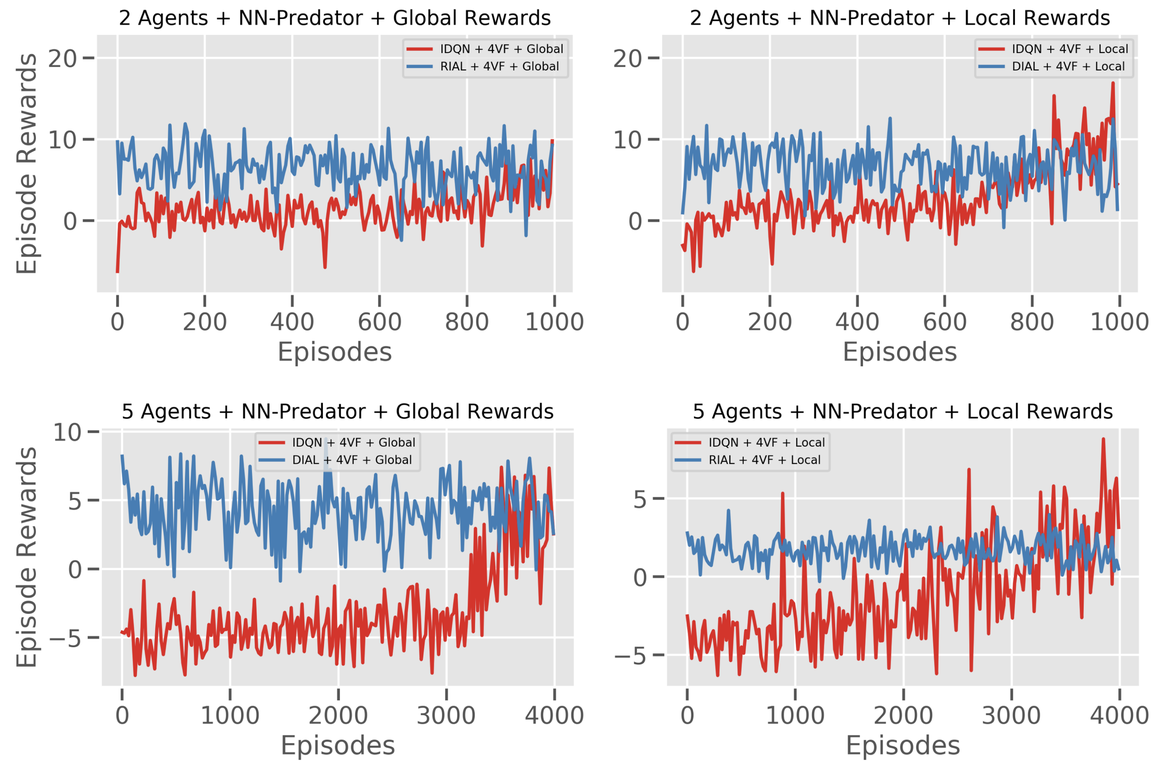


Figure 2: Learning of Swarm Behavior with Global & Local Reward Signals

- We find that local rewards enhance the learning dynamics of the agents. Furthermore, communication allows the agents to coordinate early on in learning. Later, it does not help to stabilize the non-stationary dynamics.

A Reward Curriculum for Swarm Dynamics

- We propose a curriculum learning (Bengio *et al.* , 2009) approach which successively increases the complexity of the reward signal.

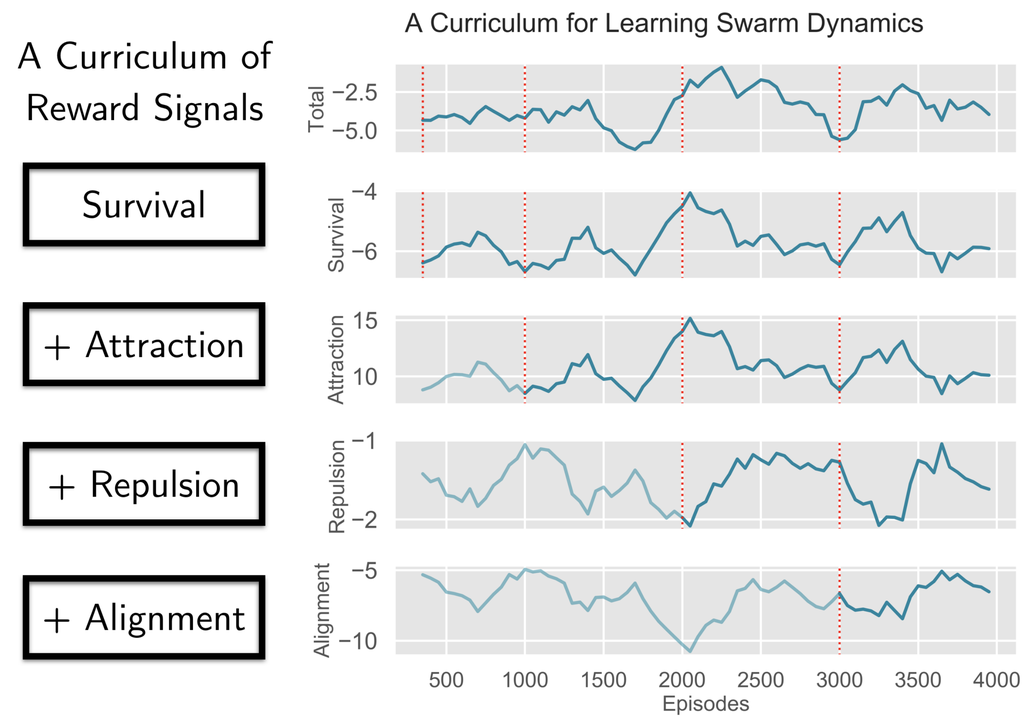


Figure 3: IDQN Learning Dynamics with a Curriculum of Rewards (5 Agents)

- Crucially, the survival and attraction as well as the repulsion and aligned objectives are coupled in their learning progress. The pairs, on the other hand, are complementary and require a behavioral trade-off.

Conclusions & Outlook

- We have introduced a **multi-agent environment to study large-scale behavior of fish schools**. Decentralized learning with global reward succeeds to result in phenomenologically plausible behavior. **Local rewards enhance learning**.
- **Communication** is crucial for rapid early progress but does not provide persistent benefits. A **curriculum of rewards** helps to analyze how reward signals are diffused by the individuals of the collective.
- Future directions: Extend environment to **shepherding behavior** as well as continuous state and action space. Extend Foerster *et al.* (2016) and differentiable communication to **locally constrained messages**.

References

Bengio, Yoshua, Louradour, Jérôme, Collobert, Ronan, & Weston, Jason. 2009. Curriculum learning. *Pages 41–48 of: Proceedings of the 26th annual international conference on machine learning*. ACM.

Foerster, Jakob, Assael, Ioannis Alexandros, de Freitas, Nando, & Whiteson, Shimon. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Pages 2137–2145 of: Advances in Neural Information Processing Systems*.

Romanczuk, Pawel, & Schimansky-Geier, Lutz. 2012. Swarming and pattern formation due to selective attraction and repulsion. *Interface focus*, **2**(6), 746–756.

Tampuu, Ardi, Matiisen, Tambet, Kodelja, Dorian, Kuzovkin, Ilya, Korjus, Kristjan, Aru, Juhan, Aru, Jaan, & Vicente, Raul. 2017. Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, **12**(4), e0172395.

GitHub Repository

rtl17@ic.ac.uk